Analysis of High Throughput Flow Cytometry Data using *plateCore*

January 2, 2009

Abstract

Background

High throughput flow studies are often run in a 96 or 384-well plate format, with a number of different samples, controls, and antibodies-dye conjugates present on each plate. Analyzing a plate requires tracking the contents of each well, matching sample wells with control wells, gating each well/channel separately, making the appropriate plots, assessing quality, and finally aggregating the results from multiple plates into an experiment level data object. This analysis can be a monumental task using traditional point-and-click software packages, even when multiple instances are deployed. We developed *plateCore* as an R/Bioconductor packaged to make processing and analysis of large, complex flow cytometry (FCM) datasets easier.

Methods

plateCore was used to analyze the results from a BD FACS™CAP screening experiment where 5 PBMC samples were assayed for 189 different human cell surface markers. This same dataset was also analyzed by a cytometry expert using FlowJo™.

Results

Positive markers identified using *plateCore* are in good agreement with those found using FlowJo™analysis.

Conclusions

plateCore provides a reproducible, objective platform for analyzing high throughput flow experiments. The R/Bioconductor implementation allows bioinformaticians and statisticians access to the data, which should further the development of automated analysis methods.

Introduction

While there are a number of different software packages available for analysis of flow cytometry data, these programs are often ill-suited to the development of new methods needed for analyzing high-throughput flow studies. Flow Cytometry High Content Screening (FC-HCS) experiments generate large volumes of data, and a systematic approach to preprocessing, gating (i.e. filtering), and summarizing results is needed for robust analyses. Ideally these steps would be automated, allowing analysis pipelines to be robust, objective, and match the high-throughput capacity of modern cytometers. Unfortunately, current approaches to FC-HCS analysis are semi-automated at best, often requiring significant manual intervention to identify cells of interest and set the appropriate gates. Since the manual contribution is subjective and prone to error when working with large numbers of samples (Maecker et al., 2005), it is desirable to develop programmatic approaches to process the data.

Flow cytometry packages available through the Bioconductor (Gentleman et al., 2004) project provide an open platform that can be used by cytometrists, bioinformaticians, and statisticians to collaboratively develop new methods for automated FC-HCS analysis. The basic data processing tools for importing, transforming, gating, and organizing raw flow cytometry data are in the *flowCore* package (Hahne et al., 2009), and the visualization functions are in *flowViz* (Sarkar et al., 2008). The Bioconductor model for flow data analysis facilitates the development of new analysis methods, since the overhead associated with accessing and visualizing flow data is handled by *flowCore* and *flowViz*. The availability of *flowCore* and *flowViz* has enabled the creation of new tools for quality assessment of large flow experiments (*flowQ* ()) and model-based clustering and automated gating (*flowClust* Lo et al. (2008)). *plateCore* also takes advantage of the functionality in *flowCore* and *flowViz* to create methods and data structures for processing large, plate-based flow datasets.

An example of the progression from raw FCM data files to a completed *plateCore* analysis is shown in Figure 1. List mode FCS files for a single plate are read into a `flowSet` using *flowCore*, and then a `flowPlate` is created by integrating the plate annotation file with the `flowSet`. The `flowPlate` is then compensated, data quality is assessed, and gates are set according to a negative control. These control gates are then applied to test wells to find cells that have specific staining in channels of interest. While this same analysis can be performed relatively quickly in other flow cytometry software packages, it can be difficult to reproduce the gating decisions made by a single expert user.

In addition to subjective gating, the lack of a standard format for describing large flow experiments also makes it difficult for anyone other than the original experimenter to replicate an analysis. The adoption of ACS specifications (ref?) should make it easier to access metadata in future flow studies, but currently this information is typically provided either as spreadsheet or a pictorial layout of a 96 well plate. Since the creation of `flowPlate` requires users to make a standard sample annotation file, plate layouts

from *plateCore* can then be easily shared along with the raw FCS2.0/3.0 files. The standard format for *plateCore* sample annotations provides a convenient way to manage the plate metadata associated with complex FC-HCS experiments.

plateCore is not designed to be a GUI driven end-user tool, but rather to help develop a standardized platform for the analysis of FC-HCS data. These analyses often represent a collaborative effort between cytometry experts who generate the data and the quantitative individuals who help deal with the large volume information. In order for this collaboration to work, the cytometrists must have confidence in the results of the automated analysis. To this point, we demonstrate the equality of our results to those produced by an expert cytometrist using FlowJo™.

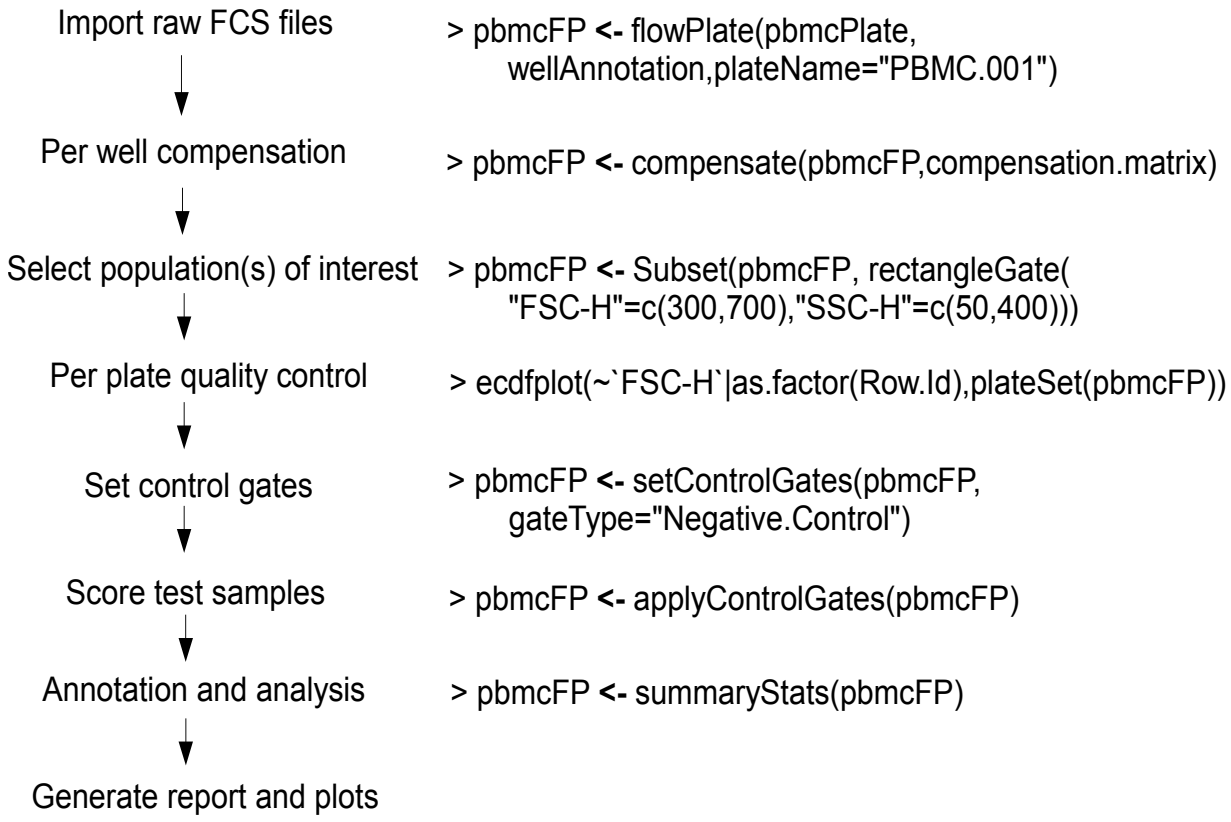

Figure 1: Typical plateCore workflow on the left, and examples of each step from a sample analysis are shown on the right. Generating reports and plots is a multi-step that typically involves merging output from several plates, and the required code is not shown here.

Materials and Methods

Data

The peripheral blood mononucleocyte (PBMC) data used in this study consists of 5 samples that were analyzed on 96-well plates using BD FACS™CAP (ref ?). On each plate, there are 189 different human cell surface antibody-dye conjugates that are arrayed 3 per well (63 test wells), along with 30 isotype control wells and 3 unstained controls. Test antibodies and isotypes are arrayed 3 per well, and the data was compensated on the cytometer (BD FACSCalibur™). The 189 antibodies were selected to provide a broad expression profile for a large number of cell surface markers, including 103 proteins with GO annotation for receptor activity, 80 for immune response, and 55 for signal transduction. The raw data is available for download from http://www.ficcs.org in the plateData.tar.gz file.

Analysis

The goal of the PBMC FACS™CAP study was to look for positive staining for the 189 different cell surface markers in lymphocytes. The *plateCore* scripts used to perform the analysis are provided in supplementary materials. Briefly, the FCM files are first processed using a combination of static (`rectangleGate`) and data driven (`norm2filter`) *flowCore* gates to pick out the lymphocytes in the forward (FSC) and side scatter (SSC) channels. The quality of the data was then assessed by looking for fluidic events such as bubbles, pressure drops, or large aggregates that can shift the baseline fluorescence readings. Fluidic events can often be identified by plotting the emprical cumulative density (ecdf) plots of FSC values for each well, and looking for distributions shifted relative to other wells (Meur et al., 2007). Based on the ecdf plots, several wells were further investigated by cytometry experts who determined that the shifts were in an acceptable range. Next the threshold between positive and negative cells are determined using the isoytpe controls, which provide a gross estimate of non-specific binding in the primary antibodies. One-dimensional gates are created using using the isotype thresholds, and these gates are applied to identify cells that are positively stained for each marker.

In addition to *plateCore*, the 5 PBMC plates were also analyzed using FlowJo™, which is one of the standard FCM data analysis platforms. First, an analysis template is created that assigns test wells and their corresponding isotype control well to a one of 30 groups. Wells in each group have similar sets of 3 antibody-dye conjugates, and the expression threshold (i.e. isotype gate) is initially set using the isotype control well. Data for each plate is imported into FlowJo™ using the template and lymphocytes were selected using a morphology (FSC-SSC) gate. Event data for the isotype well was then visualized on a log scale, and the expression threshold for each stained channel was set by picking a value that lies above the bulk of the events. For BD FACS™CAP, the isotype gate is set so that less than 1% of the events in the isotype well are above the threshold. These gates are then applied to the test wells, and the threshold may be moved up or

down based on positive test wells. The percentage of cells above the threshold for each of the 189 antibodies is then exported to a separate spreadsheet for each plate.

Results

plateCore

The 5 PBMC plates were analyzed using the approach shown in Figure 1. Results are stored in `flowPlates`, which are data structures that contain a description of the plate layout, morphology gated (FSC-SSC) events, and parameters values for the the negative control gates (i.e. isotype gates). Event level data can be visualized using plotting functions from *plateCore* and *flowViz*. Additionally, results from different plates can be aggregated, making it easier to compare results from different plates and to create the complex reports required to summarize results from 189 different markers.

Eighty-three of the 189 markers were positive on at least one plate, and these markers are shown in Figure 2. Positive for BD FACS™CAP was defined as having more than 10% of events above the isotype gate. Since antibody concentrations used in BD FACS™CAP were designed to screen a number of different cell types, the concentrations are not necessarily optimal for these PBMC samples. The 10% cutoff is an empirically determined threshold (data not shown) used to select markers for further analysis, including single color titration and competition experiments to confirm that the marker is present and staining is specific. Markers that are highly positive ($\geq$ 90%) are usually confirmed in follow-up studies, while markers that are low positives ($\leq$ 15%) are often the result of non-specific staining. Also, these percentages refer to the fraction of events above the isotype threshold, but this does not necessarily imply staining is heterogeneous.

A common goal for BD FACS™CAP screens is to identify markers that show variation in expression levels between different donors. Unfortunately, the power to detect differences in this study is limited since there are only 2 donors and the level of replication is low (2-3 plates per donor). Figure 3 shows the expected and observed variation in the percentage of positively stained cells for the 83 positive ($\geq$ 10%) markers. Histograms produced using *plateCore* are shown in Figure 4 for a selected marker, CDbd69, that exhibited a relatively high level of variation in the percentage of positive stained cells between the 5 PBMC plates.

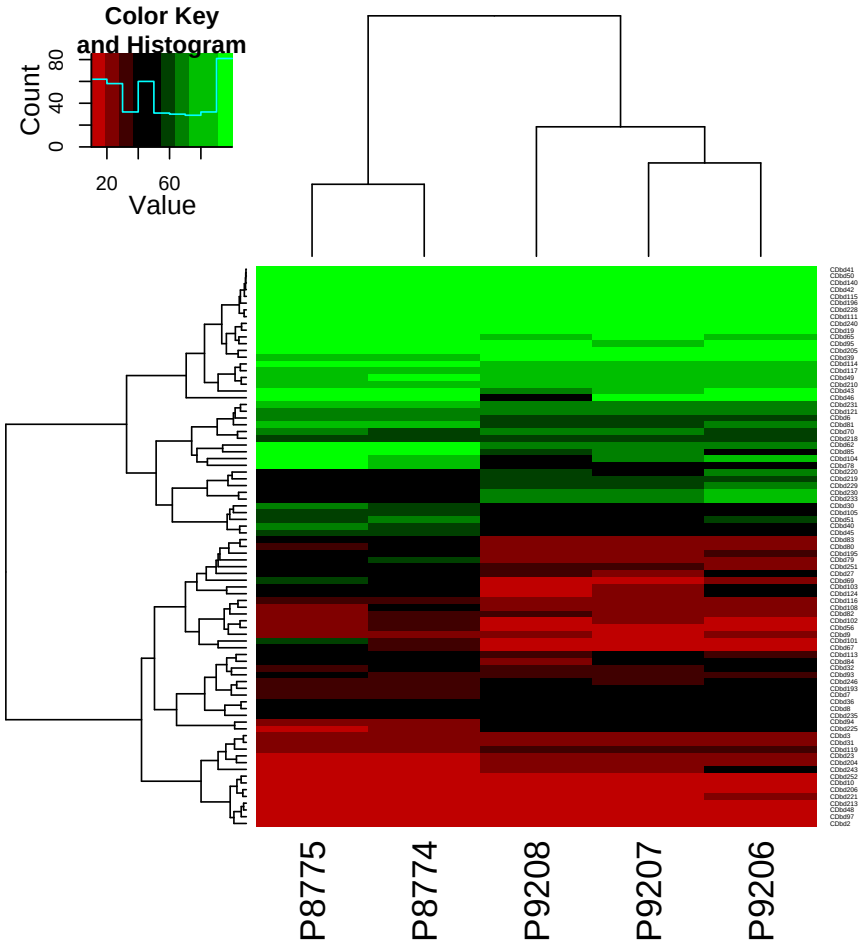

Figure 2: Heatmap showing the percentage of positive cells from the 5 different PBMC lymphocyte plates. The 83 markers with $\geq 10\%$ positive cells are shown here. Staining for the remaining 106 markers was not significantly higher than background, according to the expression threshold set using matched isotype controls. PBMCs on plates 8774-8775 and 9206-9208 are from a different individuals.

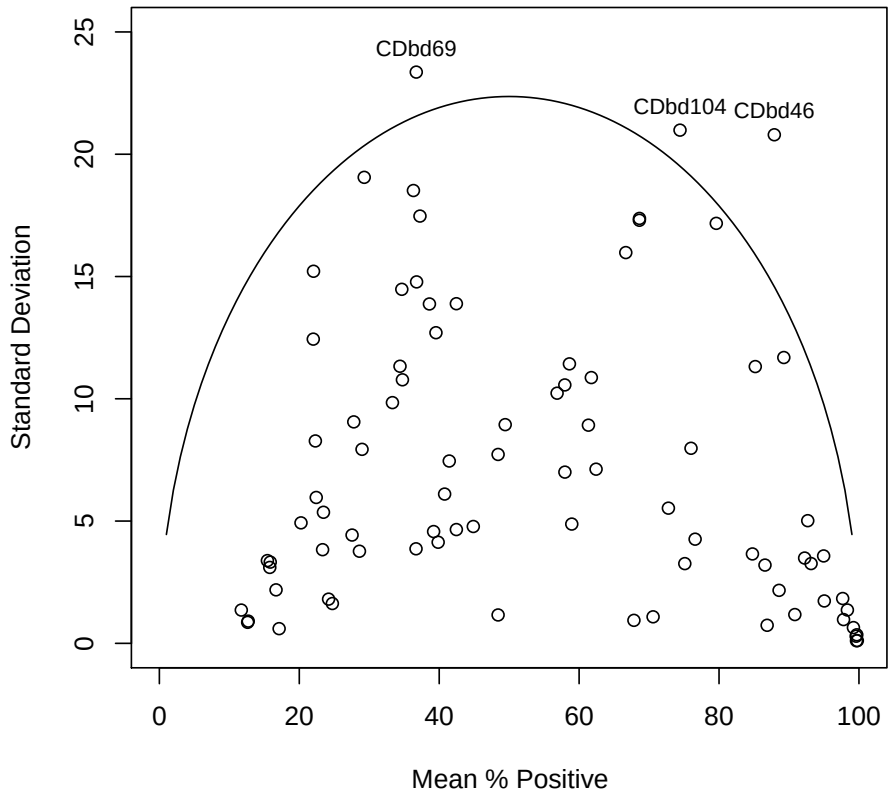

Figure 3: Scatterplot showing the mean percentage of positive cells for each positive marker versus the observed (circles) and expected (line) sample standard deviation. Expected values follow a binomial distribution with n=5. The 3 markers that lie above the expected line will be further evaluated using titration and competition experiments to see if these results represent real variation between the two donors. The curve in the expected line reflects the difficulty in gating samples whose median signal (MFI) is near the isotype cutoff, since the percentage of positive cells calculated can shift dramatically with small changes in the gate.

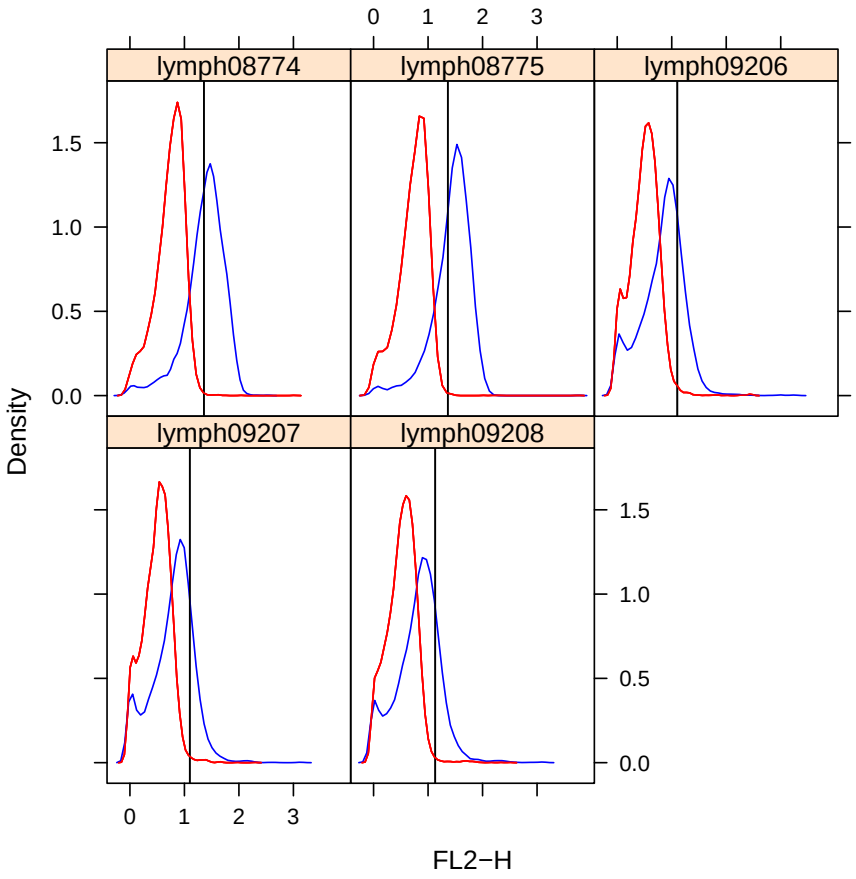

Figure 4: Histograms for CDbd69, which is one of the 3 candidates for differential expression from Figure 3. Isotypes are shown in red and test wells are in blue. Similar plots were automatically created using *plateCore* for each of 189 markers assayed in this experiment, allowing cytometry experts to quickly survey the results from the 5 different plates.

Comparison to FlowJo™Results

Automating the creation and modification of isotype gates made by cytometrists analyzing BD FACS™CAP data using FlowJo™is extremely difficult. Cytometrists adjust gates based on expert knowledge about the performance of specific antibody types and dyes, or after identifying positive test samples. The automated approach employed in *plateCore* determines the threshold using isotype controls. The gate (G_{ij}) for isotype i, channel j is set according to:

$$G_{ij} = \max(99\text{th}_{ij}, \text{MFI}_{ij} + 5\text{MAD}_{ij}), \tag{1}$$

where 99th_{ij} is the 99th percentile for the fluorescence signal, MFI is the Median Fluorescence Intensity, and MAD is Median Absolute Deviation on a linear scale. While this simple, non-parametric method works surprisingly well for BD FACS™CAP, advances in model-based clustering methods, such as those in *flowClust*, should lead to future performance improvements in automated gating. Comparisons of the *plateCore* and FlowJo™analysis are shown in Figure 5.

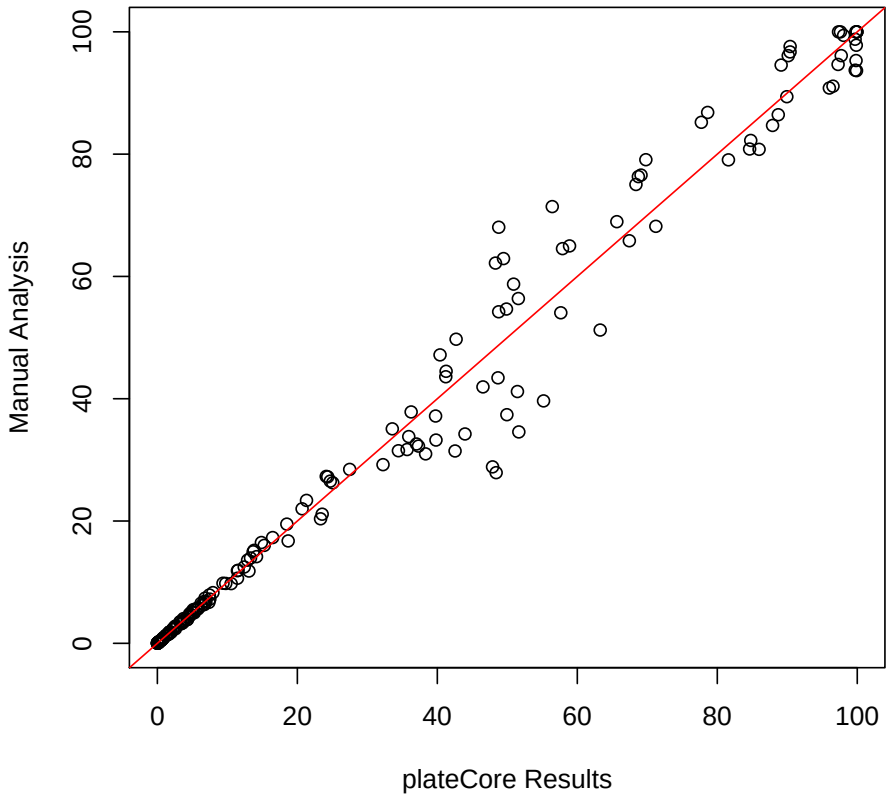

Figure 5: Percent positive results for 189 BD FACSCAP markers analyzed using either *plateCore* or manually using FlowJo™. Markers that varied the most between the the two methods tended to have intermediate percent positive values (30% to 70%), reflecting the difficulty of gating populations whose MFI is near the isotype threshold. (NOTE: Fake data, still need to get FlowJo data in this plot).

Discussion

Our approach to this PBMC BD FACS™CAP study relied on processing the raw data in parallel using both FlowJo™and *plateCore*. FlowJo™allowed the ctyometrists to thoroughly investigate individual wells, and gave them confidence that the *plateCore* results were correct (see Figure 5). Using *plateCore*, we were able to reduce the level subjectivity in setting isotype gates, eliminate mistakes associated with manual export and merging of plate output, and automate the creation of plots and data quality reports that summarized the experiment. Additionally, the *plateCore* scripts and experimental annotation can be shared with other cytometry groups, allowing them to reproduce our analysis.

The complexity of large flow experiments, like BD FACS™CAP, highlight the difficulty of applying existing flow analysis platforms to high-throughput studies. Generating and interpreting results from this PBMC study required extensive collaboration between flow cytometrists, bioinformaticians, and statisticians. At various points in the analysis, each group needed to access the raw data, annotation, and details about the experimental design. Providing this access using stand-alone flow platforms is expensive in terms of the price of multiple software licenses and in time spent training statisticians and bioinformaticians to use the programs. Fortunately the Bioconductor flow packages are modeled on standard data structures used for microarrays, which should already be familiar to most quantitative individuals working on high-throughput biological problems. We found that flowCore, *flowViz*, and *plateCore* provided an open analysis platform that facilitated communication between the flow cytometrists generating the data, and the computational experts analyzing the data.

References

Robert C Gentleman, Vincent J. Carey, Douglas M. Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J. Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y. H. Yang, and Jianhua Zhang. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004. URL `http://genomebiology.com/2004/5/10/R80`.

F. Hahne, N. LeMeur, R.R. Brinkman, B. Ellis, P. Haaland, D. Sarkar, J. Spidlen, E. Strain, and R. Gentleman. flowcore a bioconductor package for high throughput flow cytometry. *BMC Bioinformatics*, 2009.

K. Lo, RR Brinkman, and R. Gottardo. Automated gating of flow cytometry data via robust model-based clustering. *Cytometry A*, 2008.

Holden T Maecker, Aline Rinfret, Patricia D'Souza, Janice Darden, Eva Roig, Claire Landry, Peter Hayes, Josephine Birungi, Omu Anzala, Miguel Garcia, Alexandre Harari, Ian Frank, Ruth Baydo, Megan Baker, Jennifer Holbrook, Janet Ottinger, Laurie Lamoreaux, C. Lorrie Epling, Elizabeth Sinclair, Maria A Suni, Kara Punt, Sandra Calarota, Sophia El-Bahi, Gailet Alter, Hazel Maila, Ellen Kuta, Josephine Cox, Clive Gray, Marcus Altfeld, Nolwenn Nougarede, Jean Boyer, Lynda Tussey, Timothy Tobery, Barry Bredt, Mario Roederer, Richard Koup, Vernon C Maino, Kent Weinhold, Giuseppe Pantaleo, Jill Gilmour, Helen Horton, and Rafick P Sekaly. Standardization of cytokine flow cytometry assays. *BMC Immunol*, 6:13, 2005. doi: 10.1186/1471-2172-6-13. URL `http://dx.doi.org/10.1186/1471-2172-6-13`.

N. Le Meur, A. Rossini, M. Gasparetto, C. Smith, R.R. Brinkman, and R. Gentleman. Quality assessment of ungated flow cytometry data in high throughput experiments. *Cytometry A*, 71:393–403, 2007.

D. Sarkar, N. Le Meur, and R. Gentleman. Using flowViz to visualize flow cytometry data. *Bioinformatics*, 24(6):878, 2008.